EC サイトにおけるアンケートデータを考慮した購買行動分析モデルの提案

1X13C056-5　清水 良太郎
指導教員　後藤 正幸

1　研究背景・目的

近年，EC サイトを通じた購買の普及により，大量の購買履歴データを容易に蓄積できるようになった．これにより，サイト運営企業が蓄積した購買履歴データを分析し，マーケティング施策に役立てることも一般的になっている．しかし，購買履歴データのみでは,「購買傾向」は考慮できたとしても，ライフスタイルや価値観などの顧客の購買行動の背後にある意識 (以下，顧客の意識) を把握することは難しい．そこで，顧客の意識を調べるために，ユーザに対してアンケートを実施する方法がある．そして，得られた購買履歴データとアンケートデータを統合分析することで,「顧客の嗜好」を捉え，マーケティングに活用することができる．

実際に，石垣ら [1] は，Probabillistic Latent Semantic Analysis[2](以下 pLSA) と呼ばれる潜在クラスモデルをベースに，購買履歴データとアンケートデータを同時に分析し，顧客をクラスタリングするモデルを提案している．しかし，石垣らのモデルでは，購買履歴データとアンケートデータの両方が揃っている顧客のみを対象としているという制約がある．しかし，EC サイトの全ユーザに対してアンケートを行うことはコスト等の面からも困難である．実際，本研究で分析対象とするデータでは，購買履歴データに含まれるユーザが約 100,000 人であるのに対し，アンケートの調査対象で，かつアンケートに回答したユーザ (以下，回答ユーザ) は約 3,000 人のみである．そのため，石垣らのモデルを適用する場合，アンケートデータの存在しないユーザ (以下，未回答ユーザ) は分析対象とすることができない．

そこで本研究では，全ユーザの購買履歴データと，限られた少数のアンケートデータの両方を用いて，回答ユーザと未回答ユーザを顧客の意識を考慮しながら同時にクラスタリングする新たな購買行動分析モデルを提案する．このモデルを用いてクラスタリングを行うことで，顧客の意識ごとにタイプ分けすることができ，そのタイプに合わせたアイテムの推薦等，様々なマーケティング施策の立案に対する一助となる．さらに，提案モデルを服飾系 EC サイト A に蓄積された購買履歴データと，サイト A が一部のユーザに実施したアンケートデータに適用することで，その有効性を示す．

2　準備

2.1　pLSA[2]

pLSA は，ユーザとアイテムの間に潜在クラスを仮定する確率的潜在クラスモデルである．このモデルでは，潜在クラスの下でユーザとアイテムの共起が発生しているものとし，その共起関係を条件付き確率で表現する．

ここで，K 個の潜在クラス集合を $\mathcal{Z} = \{z_k : 1 \leq k \leq K\}$，$I$ 人のユーザ集合を $\mathcal{U} = \{u_i : 1 \leq i \leq I\}$，$J$ 個のアイテム集合を $\mathcal{X} = \{x_j : 1 \leq j \leq J\}$ と定義する．pLSA における完全データの確率モデルは式 (1) で表される．

$$P(u_i, x_j, z_k) \quad = \quad P(z_k)P(u_i|z_k)P(x_j|z_k) \qquad (1)$$

式 (1) の対数尤度関数を最大化するパラメータ $P(u_i|z_k)$，$P(x_j|z_k)$，$P(z_k)$ は，EM アルゴリズムによって推定する．

2.2　石垣らのモデル [1]

購買履歴データとアンケートデータを用いた購買行動分析モデルとして，石垣らのモデルがある．これは，回答ユーザのみを対象としたソフトクラスタリングに基づく潜在クラスモデルである．このモデルは，ユーザとアイテムのそれぞれに潜在クラスを仮定し，完全データの確率モデルを式 (2) で表す．ここで，ユーザの潜在クラスを $v_s(s = 1, \cdots, S)$，アイテムの潜在クラスを $w_t(t = 1, \cdots, T)$ と定義する．

$$P(u_i, x_j, v_s, w_t) \; = \; P(v_s)P(u_i|v_s)P(w_t|v_s)P(x_j|w_t) \quad (2)$$

また，アンケートデータから得られる顧客の意識に関するスコアを制約条件として，ユーザの特徴に関するパラメータ $P(v_s), P(u_i|v_s)$ を学習する．これにより，購買履歴データのみでは考慮することのできない，顧客の意識を考慮した顧客の購買行動分析が可能となる．しかし，石垣らのモデルは，全ユーザに対してアンケートデータが得られている必要がある．

3　提案手法

本研究では，全ユーザ (回答ユーザおよび未回答ユーザ) に関する購買履歴データと，回答ユーザのアンケートデータから,「購買傾向」と「顧客の意識」の両方を考慮した「顧客の嗜好」を分析するための潜在クラスモデルを提案する．

3.1　アンケートデータを用いた顧客の意識の把握

はじめに，アンケートデータから顧客の意識を分析するために，アンケートデータに対して因子分析を行う．因子分析によって多変量のアンケートデータをいくつかの因子へと集約し，各ユーザと各因子にどの程度の相関があるかを表した指標である因子得点を得ることができる．ここでは，因子数を F とし，各ユーザ u_i の因子得点ベクトルを $\boldsymbol{y}_i = (y_{i1}, y_{i2}, ..., y_{iF})^T$ と表す．

さらに，回答ユーザのアンケートデータから得られる因子得点を特徴量として，回答ユーザに関して混合正規分布モデル (以下，GMM) に基づくソフトクラスタリングを行う．GMM から得られる潜在クラス z_k の生起確率を $P_{GMM}(z_k)$，ユーザ u_i の因子得点 $\boldsymbol{y}_i$ のクラス z_k への所属確率を $P_{GMM}(z_k|\boldsymbol{y}_i)$ と表す．

3.2　提案モデルによる顧客の嗜好の把握

全ユーザの購買履歴データから得られる購買傾向と，一部のアンケートデータから得られる顧客の意識の両方を考慮した新たなクラスタリング手法を提案する．具体的には，購買履歴データを用いてクラスタリングを行う pLSA のパラメータ推定に，アンケートデータを用いて算出した回答ユーザの顧客の意識を反映するためのモデルを提案する．

本研究で提案するモデルは，式 (1) に従うものとし，EM アルゴリズムを用いてパラメータを推定する際に，GMM で学習した顧客の意識を取り込む．EM アルゴリズムにおける，E-step での $P(u_i, x_j, z_k)$ の計算において，回答ユーザについては，アンケートデータから得られる $P_{GMM}(z_k|\boldsymbol{y}_i)$ を用いる．具体的には，式 (1) における $P(u_i|z_k)$ を，以下の式 (3) ように仮定する．ここで，e_i はユーザ u_i が回答ユーザである場合に 1，それ以外では 0 をとる変数と定義する．

$$P(u_i|z_k) \; = \; \Big(\frac{P(u_i)P_{GMM}(z_k|\boldsymbol{y}_i)}{P(z_k)}\Big)^{e_i} P(u_i|z_k)^{1-e_i} \quad (3)$$

これにより，回答ユーザの各潜在クラスへの所属確率を推定する際に，購買履歴データから得られる購買傾向に併せて，アンケートデータから得られる顧客の意識を考慮することができる．また，未回答ユーザは回答ユーザのクラスタリング結果に合わせて，購買履歴データからパラメータを推定する

ことができる．すなわち，顧客の意識を考慮したパラメータ推定が可能となる．ここで，全 N 件の購買履歴データにおける n 番目のユーザを a_n，アイテムを b_n と定義する．

そして，提案モデルにおけるパラメータは，以下のようなEM アルゴリズムによって推定される．

[初期値の設定]

GMM によって得られる $P_{GMM}(z_k)$ を，$P(z_k)$ の初期値として与える．また，その他のパラメータ $P(u_i|z_k)$，$P(x_j|z_k)$ はランダムに与える．

[E-step]

$$P(z_k|u_i,x_j) = \frac{P(u_i,x_j,z_k)}{\sum_{k=1}^{K}P(u_i,x_j,z_k)} \quad (4)$$

$$P(u_i,x_j,z_k) = P(z_k)\Big(\frac{P(u_i)P_{GMM}(z_k|\boldsymbol{y}_i)}{P(z_k)}\Big)^{e_i} \times P(u_i|z_k)^{1-e_i}P(x_j|z_k) \quad (5)$$

[M-step]

$$P(z_k) = \frac{1}{N}\sum_{n=1}^{N}P(z_k|a_n,b_n) \quad (6)$$

$$P(u_i|z_k) = \frac{1}{NP(z_k)}\sum_{n=1}^{N}\delta(a_n,u_i)P(z_k|a_n,b_n) \quad (7)$$

$$P(x_j|z_k) = \frac{1}{NP(z_k)}\sum_{n=1}^{N}\delta(b_n,x_j)P(z_k|a_n,b_n) \quad (8)$$

ここで，$\delta(\alpha,\beta)$ は，$\alpha=\beta$ である場合に 1，そうでない場合には 0 の値を示すインジケータ関数である．

4　評価実験と分析

提案手法の有効性を確認するために，経営科学系研究部会連合協議会主催，平成 28 年度データ解析コンペティションで提供されたデータの大手服飾系 EC サイトの購買履歴データとアンケートデータを用いて評価実験と分析を行う．

購買履歴データの購買観察期間は 2015 年 4 月 1 日から 2016 年 3 月 31 日，総ユーザ数 I =101,501(そのうち，回答ユーザは，3,118)，総アイテム数 J =421,290，総購買回数は 1,001,901 件である．また，アンケートの全 107 項目のうち，ユーザのファッション観を考慮するために,「ファッション観に関する質問項目」の 31 問のみを使用した．また，因子数 F =12，潜在クラス数は K =5 と設定した．

4.1　評価実験

本研究の有効性を示すために，未回答ユーザがどのような顧客の意識 (ファッション観) を持つのかに関して，回答ユーザのアンケート回答データと全ユーザの購買履歴データを用いて予測することができているかを評価する．そこで，3,118 人の回答ユーザを，ランダムに Q 人のテストユーザと $3,118-Q$ 人の学習ユーザに分割する．そして，テストユーザは購買履歴データのみを，学習ユーザはアンケートデータと購買履歴データの両方を用いて，提案モデルを学習し，テストユーザの所属クラスを推定する．その結果，テストユーザのアンケートデータも用いた場合の所属クラスを正しく推定できていれば，提案モデルの有効性が示されたと判断できる．なお，今回はテストユーザの人数 Q を 1,000 人と設定した．

また，学習ユーザに対する因子分析から学習する因子負荷量と，テストユーザのアンケートデータを用いて，各テストユーザ $u_q(1\le q\le Q)$ に因子数 F 個ずつ与えられる因子得点 $\boldsymbol{y}_q^*=(y_{q1},y_{q2},...,y_{qF})^T$ を与える．さらに，この因子得点と，学習時の GMM におけるパラメータ (平均および分散共分散行列，混合比) を用いて，テストユーザのアンケートデータから得られる潜在クラスへの所属確率の正解値 $P_{GMM}(z_k|\boldsymbol{y}_q^*)$ を与える．

そして，テストユーザに関してアンケートデータから得られる所属確率の正解値と，提案手法から得られる所属確率の推定値 $P(z_k|u_q)$ の平均絶対誤差 MAE(式 (9)) を評価値とする．

$$\mathrm{MAE} = \frac{\sum_{k=1}^{K}\sum_{q=1}^{Q}\Big|P_{GMM}(z_k|\boldsymbol{y}_q^*)-P(z_k|u_q)\Big|}{KQ} \quad (9)$$

評価実験の結果，MAE の値は 0.196 であった．これに対し，$P(z_k|u_q)$ を全て 0.2 とした場合の MAE の値 0.284 をベースラインと考えれば，31%程度の改善効果が見られる．この結果より，提案モデルにおいて，パラメータの推定が有効に行われていることが示唆される．

4.2　分析結果と考察

次に，今回の提案モデルに，与えられたデータ全体を当てはめ，得られた結果に関しての分析とそれに関する考察を行う．そして，その結果は以下の通りである．

表 1．各クラスの特徴まとめ

z_k	所属しているユーザの特徴	$P(z_k)$
z_1	異性の目や周りの意見重視	0.120
z_2	流行に鈍感で，新しいものに低い関心	0.673
z_3	人の持っていないものに高い関心	0.026
z_4	伝統的なファッションブランドに精通	0.026
z_5	流行やファッションの動向に極めて敏感	0.155

表 2．各クラスに対する購買確率 Top.3 商品まとめ

	Top.1	Top.2	Top.3
z_1	ストール A	パンツ A	ブレスレット
z_2	パンツ B	シャツ/ブラウス	パンツ A
z_3	T シャツ A	ヘアゴム	スウェット
z_4	ストール A	パンツ C	T シャツ B
z_5	ストール A	カーディガン	T シャツ A

これらの結果から，クラスごとに特徴が現れていることがわかる．また，表 2 より，クラスごとに購買確率の高いと推定される商品の差異を分析することができる．クラス z_1 は「異性の目や周りの意見重視」という特徴を持っているため，他クラスのユーザも購入しているような人気商品と考えられるストール A や，パンツ A が 1 位・2 位となっている．また，クラス z_3 は「人の持っていないものに高い関心」を持っているため，他クラスには見られないようなヘアゴムやスウェットなどの商品が上位となっている．このように，購買傾向と顧客の意識の両方を考慮したクラスタリングにより，クラスごとの特徴を明らかにすることができた．

5　まとめと今後の課題

本研究では，全ユーザに関する購買履歴データから得られる「購買傾向」と，一部のアンケートデータから得られる「顧客の意識」の両方を考慮しつつ,「顧客の嗜好」を明らかにするための購買行動分析モデルを提案した．具体的には，アンケートデータが欠損しているユーザに対して，購買履歴データを用いて欠損している情報を補完し，顧客の意識を反映することのできるモデルを提案した．また，提案したモデルを実データに適用し，提案手法の有用性を示した．

今後の課題として，より正確な顧客意識の取り込み方の検討や，クラスタリングの結果に反映させる購買傾向と顧客の意識のバランスを調整できるようなモデルの考案，さらに得られた結果の活用方法を検討していくことが挙げられる．

参考文献

[1] 石垣司,“日常購買行動に関する大規模データの融合による顧客行動予測システム”, 人工知能学会論文誌 26 巻 6 号,2011.

[2] T. Hofmann,“Probabilistic Latent Semantic Analysis,” *Proc.of UAI’99*, pp.289-296, 1999.